BRIDGE: Brain-Vision Representation Integration through Depth and Granularity Encoding

Chenyuan Hong[1,*], Binghao Ye[2,*], Yufei Guo[3,†], Guoqi Li[4,5,†]

[1]School of Future Technology, South China University of Technology
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Intelligent Science & Technology Academy of CASIC
[4]Institute of Automation, Chinese Academy of Sciences
[5]Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology

Abstract

Decoding visual content from non-invasive brain signals remains challenging because neural responses evolve over time whereas images are typically static. Most existing methods align an entire neural response window to a single final visual embedding, potentially overlooking a fundamental representational mismatch: brain signals reflect both early perceptual and later integrative representations, whereas the final visual embedding is biased toward high-level semantics. Therefore, we propose BRIDGE, a brain-vision representation integration framework that aligns two modalities through Visual Depth Encoding and Brain Granularity Encoding. On the visual side, BRIDGE extracts and fuses CLIP representations from multiple depths, producing an alignment target that preserves low-level and high-level information. On the brain side, BRIDGE partitions stimulus-evoked EEG/MEG responses into a small number of ordered temporal segments and aggregates their representations through learnable attention pooling, allowing the model to capture temporal heterogeneity. The resulting brain and visual embeddings are trained in a shared latent space by contrastive learning and can further support brain-to-image generation through a pretrained diffusion prior. Experiments on THINGS-EEG and THINGS-MEG demonstrate that BRIDGE achieves strong retrieval and generation performance. Ablation studies further confirm the complementary benefits of depth-wise visual aggregation and temporal brain factorization. Code is available at `https://github.com/hcy1026/BRIDGE`.

1 Introduction

Decoding visual content from non-invasive brain signals [1, 2] has become a rapidly advancing problem at the intersection of machine learning, neuroscience, and brain-computer interfaces [3, 4]. Within this field, driven by recent progress in multimodal representation learning and generative modeling, a growing line of work learns a shared latent space between brain signals and visual representations, enabling classification, retrieval, and reconstruction directly from neural recordings [5–13].

Existing methods have largely advanced along an alignment-centered trajectory. Early approaches, such as NICE [5], BraVL [6], and ATM [7], map brain signals into pretrained visual embedding spaces through contrastive learning, multimodal generative learning, or CLIP-guided projection. Subsequent methods, including MB2C [9], UBP [10], Neural-MCRL [8], and NeuroBridge [11], extend this trajectory by introducing cycle consistency, uncertainty-aware priors, semantic bridging,

*Equal contribution.
†Corresponding authors. Email: yfguo@pku.edu.cn, guoqi.li@ia.ac.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and bidirectional cross-modal constraints. Collectively, these studies have established brain-signal-based visual decoding as a highly active research direction.

Nevertheless, existing decoding pipelines largely overlook a fundamental representational mismatch between brain and visual embeddings. Visual information in the brain is not encoded as an instantaneous static state, but unfolds through temporally evolving stimulus-evoked responses. Time-resolved decoding studies show that early neural responses are primarily driven by perceptual attributes, such as contours, spatial structure, whereas later responses increasingly reflect integrative information related to object category and conceptual meaning [14, 15]. This temporal organization suggests that brain signals contain a mixture of perceptual and integrative evidence, which will be treated equally by brain encoder networks. In contrast, images are static. Prior studies from Cichy et al. [16], Radford et al. [17] show that CLIP-like and related visual backbones exhibit a hierarchical representational structure, where shallow layers preserve low-level texture and structure, while deeper layers capture higher-level semantic contents. However, most existing neural decoding pipelines rely only on the final visual embedding, thereby introducing a mismatch: brain activity reflects early perceptual low-level and later integrative high-level representations, whereas the final visual embedding is biased toward deep high-level semantic abstraction. Although BrainHiVE [18] incorporates multiple pretrained encoders, it still treats the representational hierarchy as a set of discrete fusion endpoints, rather than modeling the continuous, overlapping, and sustained dynamics of visual processing [8, 10, 14, 16].

From this perspective, we propose BRIDGE, as illustrated in Fig. 1, a better alignment framework for Brain-Vision Representation Integration through Depth and Granularity Encoding. Rather than aligning brain signals to a single endpoint, high-level visual embedding, BRIDGE integrates temporally staged neural information and multi-depth visual information into a shared embedding for multilevel alignment. On the visual side, we introduce Visual Depth Encoder (VDE), which aggregates image representations from multiple depths of the visual encoder, preserving both low-level cues and high-level semantics. On the brain side, to extract brain signals more adaptively across the temporal hierarchy, rather than representing the entire response window as a single undifferentiated input, we introduce the Brain Granularity Encoder (BGE), which partitions the stimulus-evoked response window into a small number of temporally ordered segments and adaptively integrates them. By integrating temporally structured neural representations with multi-depth visual features, BRIDGE provides a neuroscience-inspired and information-preserving framework for visual decoding from brain signals.

In summary, our contributions are four-fold:

- We identify a hierarchical representation mismatch between temporally evolving brain-signal representations and static late-stage visual embeddings, and propose a new **BRIDGE** framework based on this insight.
- We introduce **Brain Granularity Encoder (BGE)**, which organizes the stimulus-evoked brain response window into a small number of temporally ordered processing stages, enabling stage-specific neural information to be modeled without treating the full response window as homogeneous.
- We introduce **Visual Depth Encoder (VDE)**, which aggregates visual representations along the depth of a pretrained visual encoder, integrating low-level cues and high-level semantic information into a unified alignment target.
- Extensive experiments demonstrate that BRIDGE achieves competitive performance on brain-signal-based visual decoding benchmarks.

2 Related works

Brain-signal-based visual decoding. Visual decoding from non-invasive brain signals has progressed from closed-set classification toward representation-level alignment and reconstruction. NICE establishes contrastive brain-image alignment for zero-shot object decoding [5]. BraVL proposed by Du et al. [6] introduces a brain-visual-linguistic latent model to exploit textual semantics under limited paired neural data [6]. ATM maps brain signals into a CLIP-aligned space and further uses diffusion priors for reconstruction [7]. In parallel, fMRI-based methods such as MindEye from Scotti et al. [19], Brain-Diffuser from Ozcelik and VanRullen [20], and latent-diffusion reconstruction from Takagi

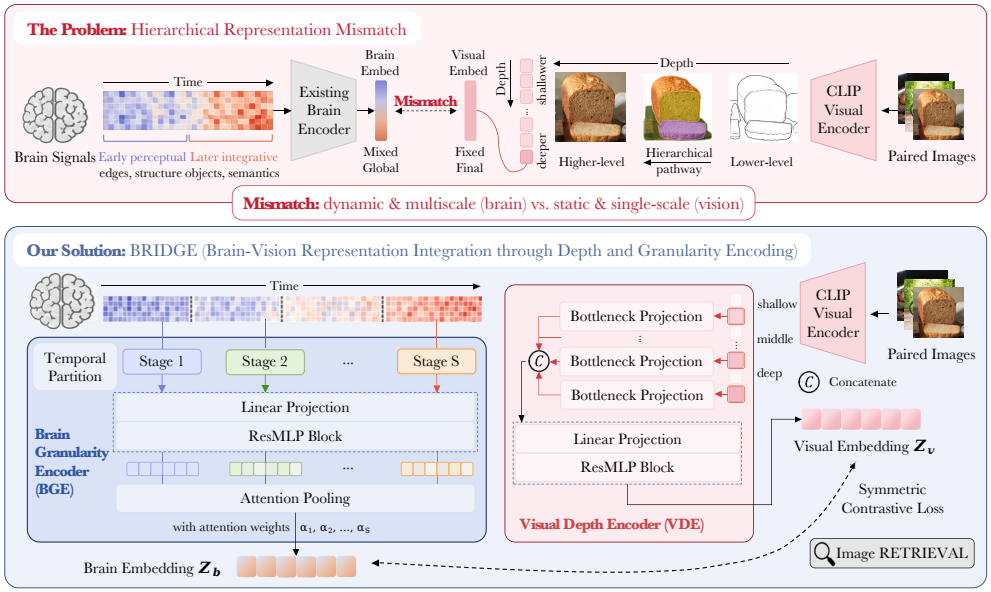


Figure 1: Overview of the proposed BRIDGE framework. The upper panel illustrates the hierarchical representation mismatch between brain signals and visual representations. The lower panel shows our solution, which aligns these two modalities through Brain Granularity Encoder (BGE) and Visual Depth Encoder (VDE). The resulting brain embedding Z_b and visual embedding Z_v are aligned in a shared latent space using a symmetric contrastive loss.

and Nishimoto [21] show that generative priors can recover image content from brain activity. More recent brain-signal methods pay attention to the EEG/MEG data itself, and focus on reducing modality mismatch by imposing bidirectional cycle consistency [9], modeling systematic and random gaps with an uncertainty-aware blur prior [10], introducing semantic bridging and cross-attention [8], using cognitive prior augmentation with bidirectional semantic alignment [11] and leveraging multiple pretrained visual encoders with different inductive biases [18]. However, these methods typically rely on a final visual embedding, a generative prior, or externally assembled visual targets, leaving the internal layer-wise semantic trajectory of pretrained image encoders under-exploited for brain-visual multi-stage alignment.

Layer-wise semantics in visual foundation models. Deep visual encoders organize image information unevenly across depth. Classical visualization and dissection studies show that lower layers encode local edges, textures, colors, and materials, while deeper layers increasingly correspond to parts, objects, scenes, and semantic concepts [22, 23]. Task-optimized hierarchical models also predict neural responses along the ventral visual stream, suggesting that model depth carries meaningful representational structure rather than redundant computation [24, 25]. CLIP further provides a strong vision-language embedding space trained from large-scale image-text supervision [17], but its final embedding is optimized for semantic transfer rather than preserving all visual details. Analyses of ViT and CLIP-ViT show that transformer layers preserve different degrees of spatial, structural, and semantic information, and that final layers often behave as global semantic aggregation stages [26–29]. Therefore, using only the final CLIP embedding as the alignment target may discard intermediate visual evidence that remains relevant to brain-signal decoding.

Time-resolved visual representations in brain signals. Visual representations in the brain are temporally organized rather than static. MEG studies show that object information emerges within the first few hundred milliseconds after stimulus onset, with time-resolved patterns supporting both exemplar-level and category-level decoding [14, 30, 31]. Comparisons between deep networks and cortical dynamics further indicate a hierarchical correspondence: earlier neural responses are better explained by shallower visual representations, whereas later responses increasingly match higher-level model features [16, 32]. Scene-decoding studies by Martin Cichy et al. [33] similarly show that different visual properties, such as layout and scene size, arise at distinct latencies in MEG

responses. Recent intracortical evidence from Shi et al. [34] further demonstrates that macaque inferotemporal face patches use temporally distinct coding stages, shifting from early general coding to later face-specific coding. Large-scale EEG datasets for object recognition also confirm that high-temporal-resolution non-invasive recordings contain rich stimulus-related dynamics suitable for computational modeling [35]. These findings motivate treating the brain-signal response window as temporally heterogeneous, while the implementation should avoid fragmenting low-SNR neural evidence into unstable fine-grained segments.

3 Method

3.1 Overview and problem statement

Our work proposes a *Brain-Vision Representation Integration through Depth and Granularity Encoding* framework for brain-signal-based visual decoding as illustrated in Fig. 1, whose key idea is to align multi-stage brain and visual embeddings. While preserving their distinct structural properties, brain signals are naturally multiscale in *time*, whereas visual representations are in *network depth*. We therefore design Brain Granularity Encoder (BGE) modeling brain signals with temporal segmentation and segment-wise attention pooling, and Visual Depth Encoder (VDE) modeling images with multi-depth CLIP features followed by trainable branch fusion. The two modalities are then aligned in a shared embedding space with a symmetric contrastive objective.

Formally, let

$$\mathcal{D} = \{(\mathcal{X}_i, \mathcal{I}_i, s_i)\}_{i=1}^{N}$$

be a paired dataset, where $\mathcal{X}_i \in \mathbb{R}^{C\times T}$ represents the i-th brain signal sample with T indicating the length of the time window, C the number of electrode channels, $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$ is the corresponding visual stimulus with H and W denoting the image height and width respectively, N is the total number of data samples and s_i is the subject index. We learn a brain encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and a visual encoder $g_\phi : \mathcal{I} \to \mathbb{R}^d$ to project brain signals and images into a shared latent space, where the corresponding embeddings Z_b and Z_v are aligned by optimizing a contrastive learning objective.

3.2 Brain Granularity Encoder (BGE)

The stimulus-evoked brain response unfolds over time, with early segments carrying predominantly perceptual information and later segments reflecting increasingly integrative, semantic content. To adaptively capture this temporal hierarchy without treating the full response window as homogeneous, we propose the BGE, which partitions the temporal axis into ordered stages and learns to dynamically emphasize the most informative segments.

Temporal partition. Given $\mathcal{X}_i \in \mathbb{R}^{C\times T}$, we uniformly partition its temporal axis into S ordered segments (with $S = 2$ by default, i.e., a two-segment temporal partition). Let

$$0 = t_0 < t_1 < \cdots < t_S = T$$

denote the partition boundaries, defined by

$$t_m = \begin{cases} m\left\lfloor \dfrac{T}{S} \right\rfloor, & m = 0, 1, \ldots, S-1, \\ T, & m = S. \end{cases} \tag{1}$$

The first $S-1$ segments have equal length, while the final segment absorbs the remainder when T is not divisible by S. The m-th segment is defined as

$$\mathcal{X}_i^{(m)} = \mathcal{X}_i[:, t_{m-1} : t_m] \in \mathbb{R}^{C\times T_m}, \qquad T_m = t_m - t_{m-1}, \tag{2}$$

where $m \in \{1, \ldots, S\}$, and T_m denotes the length of the m-th temporal segment. To enable parameter sharing across segments of unequal lengths, we zero-pad each segment to

$$T_{\max} = \max_{m=1,\ldots,S} T_m. \tag{3}$$

Shared segment encoder. For the m-th segment of sample i, we apply zero-padding and flattening:

$$u_i^{(m)} = \mathrm{Flatten}\Big(\mathrm{Pad}\Big(\mathcal{X}_i^{(m)}\Big)\Big) \in \mathbb{R}^{CT_{\max}}, \tag{4}$$

where $\mathrm{Pad}(\cdot)$ denotes zero-padding along the temporal axis, $\mathrm{Flatten}(\cdot)$ reshapes the resulting matrix into a vector in $\mathbb{R}^{CT_{\max}}$.

We then project $u_i^{(m)}$ into the shared d-dimensional embedding space using a learned linear map W_b:

$$\bar{h}_i^{(m)} = u_i^{(m)} W_b, \qquad W_b \in \mathbb{R}^{CT_{\max} \times d}, \tag{5}$$

To further refine the projected feature, we apply a Residual Multi-Layer Perceptron (MLP) block:

$$h_i^{(m)} = \mathrm{LayerNorm}\Big(\bar{h}_i^{(m)} + \phi_b\Big(\bar{h}_i^{(m)}\Big)\Big) \in \mathbb{R}^d, \tag{6}$$

where $\phi_b : \mathbb{R}^d \to \mathbb{R}^d$ denotes a two-layer MLP with a hidden dimension d_s and GELU activation. The resulting vector $h_i^{(m)}$ is the embedding of the m-th temporal segment of sample i.

MLP-style attention pooling. Given the segment embeddings $\{h_i^{(m)}\}_{m=1}^S$ for sample i, we aggregate them into a single brain representation using a learnable MLP-style attention pooler. Specifically, we assign each segment a scalar importance score

$$a_i^{(m)} = \rho_\omega\Big(h_i^{(m)}\Big), \tag{7}$$

where $\rho_\omega : \mathbb{R}^d \to \mathbb{R}$ is a learnable scoring function parameterized by ω, a two-layer MLP scorer with a $\tanh$ activation, shared across all segments and samples. We then normalize these scores across the S segments of the same sample via a softmax operation:

$$\alpha_i^{(m)} = \frac{\exp(a_i^{(m)})}{\sum_{r=1}^S \exp(a_i^{(r)})}, \tag{8}$$

where $\alpha_i^{(m)} \in (0, 1)$ is the normalized attention weight of segment m for sample i, satisfying $\sum_{m=1}^S \alpha_i^{(m)} = 1$. The final brain embedding z_i^b is defined as the weighted combination

$$z_i^b = \sum_{m=1}^S \alpha_i^{(m)} h_i^{(m)} \in \mathbb{R}^d, \tag{9}$$

The final representation is a learned weighted combination of shared segment embeddings, allowing the model to emphasize temporally informative segments instead of treating all segments uniformly.

Brain signals often exhibit substantial subject-dependent static offsets, especially in inter-subject settings [36, 37]. To reduce this nuisance variation, we optionally apply baseline subtraction before the temporal partition, the details are in Appendix C.

3.3 Visual Depth Encoder (VDE)

Visual representations in deep networks are organized hierarchically across depth, with shallow layers preserving low-level structure and deeper layers capturing high-level semantics. To preserve this multi-scale visual information rather than discarding intermediate layers, and to provide a hierarchically compatible alignment target for the temporally staged brain representations, we propose the VDE, which aggregates features from multiple depths of a pretrained CLIP encoder into a unified embedding that can be directly matched with the brain-side multi-stage outputs.

Multi-depth CLIP features. For the visual modality, we adopt a pretrained CLIP image encoder as the frozen visual backbone. Given an input image $\mathcal{I}_i$, the CLIP [17, 38] encoder produces hidden representations at multiple depths and the final embedding. Let $\mathcal{L}$ denote the set of selected transformer depths, and let $|\mathcal{L}|$ denote the number of selected branches. For each depth $\ell \in \mathcal{L}$, we extract a visual feature

$$v_i^{(\ell)} = E_{\mathrm{CLIP}}^{(\ell)}(I_i), \qquad \ell \in \mathcal{L}, \tag{10}$$

where $E_{\mathrm{CLIP}}^{(\ell)}(\cdot)$ denotes the hidden representation at depth ℓ, and $v_i^{(\ell)}$ is the corresponding feature for image $\mathcal{I}_i$. The selected depths span a progression from relatively lower-level pixel representations to higher-level semantic representations.

Branch-specific bottleneck projection. The raw CLIP features extracted from different depths may differ substantially in feature distribution and dimensions, so we project each branch independently into a shared feature space of equal dimensionality using a branch-specific dimensionality reduction map before fusion:

$$b_i^{(\ell)} = W_\ell v_i^{(\ell)} + c_\ell \in \mathbb{R}^{d_b}, \tag{11}$$

where $v_i^{(\ell)}$ is the CLIP feature at depth ℓ for image I_i, W_ℓ and c_ℓ are trainable branch-specific parameters and $d_b = 256$ in our setting. This projection not only improves cross-depth compatibility, but also reduces the parameter cost and acts as a learnable information bottleneck that preserves task-relevant components.

Cross-depth fusion. After branch-wise projection, we concatenate the projected representations from all selected depths to form a unified multi-depth visual descriptor:

$$b_i = \mathrm{Concat}\big(b_i^{(\ell)}\big)_{\ell\in\mathcal{L}} \in \mathbb{R}^{|\mathcal{L}|d_b}, \tag{12}$$

where $\mathrm{Concat}(\cdot)$ denotes concatenation along the feature dimension. We then map this concatenated descriptor into the shared latent space through a Residual MLP block:

$$\bar{z}_i^v = b_i W_v, \qquad W_v \in \mathbb{R}^{|\mathcal{L}|d_b \times d} \tag{13}$$

$$z_i^v = \mathrm{LayerNorm}(\bar{z}_i^v + \phi_v(\bar{z}_i^v)) \in \mathbb{R}^d, \tag{14}$$

where $\phi_v : \mathbb{R}^d \to \mathbb{R}^d$ denotes a two-layer MLP with GELU activation, and z_i^v is the final visual embedding for sample i.

Throughout training, the CLIP backbone E_{CLIP} remains frozen, and only the branch projection parameters $\{W_\ell, c_\ell\}_{\ell\in\mathcal{L}}$ and the MLP parameters $\{W_v, \phi_v\}$ are optimized. This design preserves the strong visual prior provided by pretrained CLIP while allowing the target embedding space to adapt to cross-modal alignment with brain signals.

3.4 Symmetric brain-vision contrastive alignment

We align the brain embedding z_i^b and the visual embedding z_i^v with a symmetric CLIP-style InfoNCE loss followed by van den Oord et al. [39]. First, we apply L2 Normalization on both modalities to obtain $\hat{z}_i^b$ and $\hat{z}_i^v$. Given a minibatch of size B, the cosine-similarity logits are

$$s_{ij} = \exp(\gamma)\,(\hat{z}_i^b)^\top \hat{z}_j^v, \tag{15}$$

where γ is a learnable logit-scale parameter (initialized to $log(1/0.07)$). The final objective is

$$\mathcal{L}oss = -\frac{1}{2B}\left(\sum_{i=1}^{B}\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \sum_{i=1}^{B}\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right). \tag{16}$$

The training procedure is shown in Appendix E.

3.5 Extension to Brain-to-Image Generation

We follow the generation architecture of BrainHiVE [18] and interface the learned brain representation with a pretrained diffusion prior via a lightweight conditioning pathway. The brain signal is encoded by Brain Granularity Encoder (BGE), whereas the visual stimulus is encoded by Visual Depth Encoder (VDE), which serves as the target representation space for generation-oriented alignment. We first pretrain a vision-to-diffusion prior that maps VDE features to the conditioning space of a frozen diffusion generator. Given a VDE embedding Z_v, a residual MLP block transforms it into a diffusion condition, which is injected into a frozen SDXL UNet [40] through a training IP-Adapter [41]. After pretraining, both the diffusion backbone and the visual-side prior are kept frozen, and only the BGE is optimized using paired brain-image samples. During inference, the BGE encodes the input brain signal, and the resulting representation is used by the pretrained conditioning pathway to generate images. Additional details are provided in Appendix D.

4 Experiments

4.1 Datasets and experimental settings

Datasets We evaluate our method on two large-scale benchmarks for visual decoding from non-invasive brain signals: **THINGS-EEG** and **THINGS-MEG**. For THINGS-EEG [35, 42], the benchmark comprises 10 participants, with 1,654 training concepts and 200 test concepts. We further evaluate BRIDGE on THINGS-MEG to assess its applicability across different neural recording modalities. THINGS-MEG [42] contains recordings from 4 participants and covers 1,654 training concepts and 200 test concepts from the THINGS database. To improve the signal-to-noise ratio, repeated trials of the same stimulus are averaged within each subject before evaluation in both datasets. In THINGS-EEG, each image is presented for 100 ms, followed by a 100-ms blank interval. In THINGS-MEG, each image is presented for 500 ms, followed by a fixation interval of 1000 ± 200 ms. For both datasets, BRIDGE uses the 0–1000 ms post-stimulus response window, excluding pre-stimulus samples.

Experimental settings We follow the standard zero-shot evaluation protocol and data splits adopted in recent prior work Zheng et al. [18]. By default, we set the number of temporal segments to 2, use LAION CLIP ViT-B/32 [43] as the visual encoder, and select 4 visual depths. All experiments are conducted on a single NVIDIA 4090 24 GB GPU. Additional dataset and implementation details are provided in the Appendix A.

4.2 Comparison with State-of-the-Art Methods

We evaluate our model on both brain-image retrieval and brain-to-image generation.

For retrieval, we evaluate performance using 200-way zero-shot top-1 and top-5 accuracy on THINGS-EEG and THINGS-MEG, under both intra-subject and inter-subject evaluation protocols. Table 1 compares our method with representative prior approaches, including BraVL [6], NICE [5], ATM [7], CognitionCapturer (C.C.) [44], UBP [10] and other methods. On THINGS-EEG, our method achieves **80.8%|97.0%** Top-1|Top-5 accuracy under the intra-subject protocol, surpassing the strongest baseline by 5.1 and 2.4 percentage points, respectively. Under the inter-subject protocol, it reaches **33.7%|65.8%**, compared with **20.0%|44.1%** reported by Zheng et al. [18]. On THINGS-MEG, our method obtains competitive performance across both protocols, achieving **33.1%|61.5%** in the intra-subject setting and **7.4%|19.3%** in the inter-subject setting.

For generation, we evaluate the quality of neural image reconstruction on both EEG and MEG, as summarized in Table 13. We compare BRIDGE with representative reconstruction baselines, including B.D. [45], ATM [7], AVDE [46], C.C.(All) [44], SeeEEG [47], and BrainHiVE [18]. Following prior works, we assess reconstruction quality using both low-level and high-level metrics. Specifically, low-level quality is measured by PixCorr and SSIM. High-level semantic consistency is evaluated by two-way identification [48] using AlexNet(2), AlexNet(5), Inception, and CLIP, where higher values indicate better performance. For SwAV, lower values indicate better performance. As shown in Fig. 2 , we reconstruct images from EEG recorded from different subjects. Additional quantitative results and comparisons are reported in Appendix F.2.

The overall performance remains comparable to state-of-the-art methods, indicating that our approach transfers reasonably well across modalities and is effective for both retrieval and generation.

4.3 Ablation study

To better understand where the performance gains come from, we conduct a series of ablation studies on the multi-stage design from both sides of the alignment framework. We first compare the full model with variants that remove the multi-stage design from the brain branch, the image branch, or both. This analysis allows us to isolate the contribution of each component, Table 2 demonstrates that BGE and VDE provide complementary improvements.

Given that the brain-side improvement comes from explicitly modeling temporal heterogeneity, it is important to determine the appropriate temporal granularity. Neuroscience does not suggest that visual processing should be partitioned into a fixed number of discrete stages. Instead, it supports a

Table 1: Comparison of 200-way zero-shot retrieval accuracy on the THINGS-EEG and THINGS-MEG benchmarks. Results are reported as subject-wise averaged Top-1 and Top-5 accuracies (%) under both intra-subject and inter-subject evaluation protocols. "–" indicates that the corresponding result is not reported.

Method	EEG				MEG			
	Intra-subject		Inter-subject		Intra-subject		Inter-subject	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
BraVL [6]	5.8	17.5	1.8	7.0	–	–	–	–
NICE [5]	16.1	43.6	6.2	21.4	12.8	36.0	–	–
ATM [7]	27.1	58.1	11.9	33.8	–	–	–	–
VE-SDN [49]	37.2	69.9	–	–	–	–	–	–
CC-All [44]	35.6	80.2	–	–	–	–	–	–
Neural-MCRL [8]	32.4	64.1	14.0	34.3	–	–	–	–
UBP [10]	50.9	79.7	12.4	33.4	26.7	55.2	2.2	10.4
SRT [47]	26.3	59.5	13.8	37.1	–	–	–	–
NeuroBridge [11]	63.2	89.9	19.0	45.9	–	–	–	–
AVDE [46]	30.0	58.2	14.3	32.9	23.6	49.8	–	–
BrainHiVE [18]	75.7	94.6	20.0	44.1	**33.7**	60.5	5.4	15.2
BRIDGE(Ours)	**80.8**	**97.0**	**33.7**	**65.8**	33.1	**61.5**	**7.4**	**19.3**

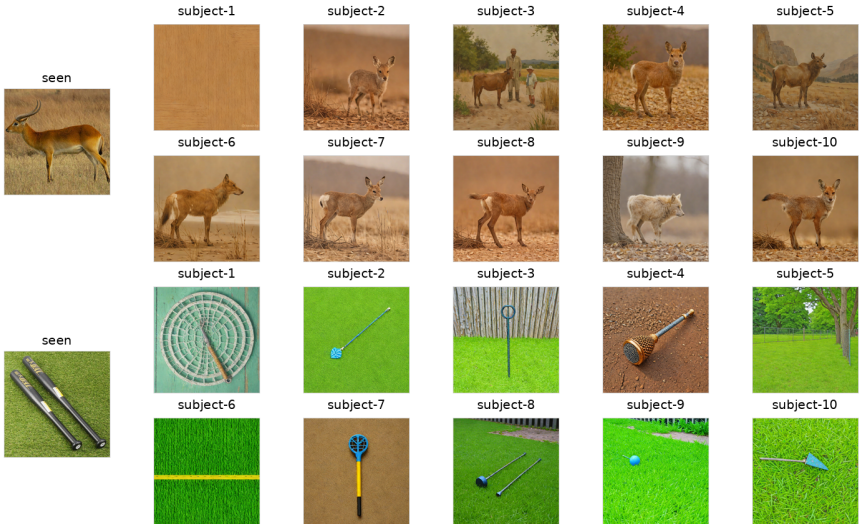


Figure 2: Comparison between seen images and reconstructed images from subject-1 to subject-10.

coarse temporal transition from earlier perceptual analysis to later semantic processing. Therefore, we treat the number of temporal stages S as a modeling granularity rather than a biological constant.

Our ablation follows the baseline setting of Zheng et al. [18], using three pretrained visual encoders and a whole-temporal MLP-based brain encoder. Taking EEG as an example, Table 3 (a) shows that splitting EEG into two temporal stages substantially improves performance over a single global representation. However, further uniform partitioning into three, four, or five stages degrades performance, while the six-stage setting only partially recovers the accuracy. We interpret the two-stage design as an effective architectural approximation of the transition

Table 2: Joint ablation results under different settings. Results are reported as subject-wise averaged Top-1 and Top-5 accuracies (%) under intra-subject evaluation protocol.

Setting	Top-1 (%)	Top-5 (%)
w/o BGE	77.10	96.10
w/o VDE	51.85	82.72
w/o both	48.98	80.90
FULL	**80.75**	**97.03**

in visual neural responses. Compared with a whole-window representation, it exposes temporal processing stages that are useful for brain-image alignment; compared with finer partitions, it preserves sufficient signal within each segment from low-SNR and non-stationary EEG responses for stable cross-modal matching [50]. Furthermore, recent intracortical evidence provides a concrete example of two-stage transition: from initially general object-like coding to a face-specific coding supporting identity discrimination around 100 ms [34]. Therefore, we adopt $S = 2$ as a simple, biologically motivated, and empirically robust temporal factorization for brain-vision alignment in our setting.

Table 3: Ablation studies on the number of EEG stages and image stages. Results are reported as subject-wise averaged Top-1 and Top-5 accuracies (%) under the intra-subject evaluation protocol.

(a) EEG-stage ablation

EEG stages	Top-1	Top-5
Baseline	75.70	94.60
2	**78.87**	**96.15**
3	75.92	94.58
4	66.33	91.33
5	33.52	60.63
6	46.65	77.67

(b) Image-stage ablation

Intermediate image stages	Top-1	Top-5
Baseline	75.70	94.60
2 (5, 11)	73.23	94.50
3 (3, 7, 11)	75.73	95.32
4 (2, 5, 8, 11)	77.13	96.10
5 (2, 4, 6, 8, 11)	77.85	96.10
6 (1, 3, 5, 7, 9, 11)	79.10	96.65

We further provide clustering visualizations in Appendix G, which offer additional evidence that two temporal segments yield the most reasonable and robust design.

After analyzing temporal factorization on the brain side, we further investigate the visual-side multi-stage design. We utilize uniformly-spaced sampling features from multiple depths (1- 12 [17, 43]) allowing the model to gain a richer representational trajectory rather than a single endpoint. To examine whether the choice is well justified, we compare several layer combinations and study how different depth distributions influence the performance. Each setting contains the $|\mathcal{L} - 1|$ intermediate layer outputs and the final embedding, we ignore the 12th layer since it is similar to the final. The results are summarized in Table 3 (b). We treat intermediate CLIP layers as representations at different depths and evaluate several distributed layer subsets. We find that multiple evenly distributed selections all perform strongly, (2,5,8,11), (2,4,6,8,11) and (1,3,5,7,9,11) share almost identical high performances. This suggests that the benefit comes primarily from covering the entire representational depth of the visual encoder, rather than from any single privileged layer combination or from the number of stages. Therefore, we select layers (2,5,8,11) by default as a trade-off between representation coverage and computational efficiency.

Table 4: Joint summary of image-stage and EEG-stage ablations.

EEG stages	Image stages	Top-1 (%)	Top-5 (%)
baseline	baseline	75.70	94.60
2	6	82.58	97.06
4	4	69.28	92.88
6	4	48.97	80.88
2	2	77.85	96.00
4	2	66.65	91.29

To further verify the effect of jointly modeling temporal stages on both the brain and visual sides, we report a combined stage ablation in Table 4.

5 Discussion and conclusion

We proposed a *Brain-Vision Representation Integration through Depth and Granularity Encoding* framework for brain-visual alignment. The main motivation is the fundamental mismatch between brain and visual embeddings. On the brain side, Brain Granularity Encoder partitions the response window and adaptively aggregates temporally informative components, making the mixed perceptual and semantic evidence in brain signals more explicit. On the visual side, Visual Depth Encoder extracts and fuses multiple CLIP layers, replacing a single final-layer semantic endpoint with a richer target that spans multiple levels of visual abstraction. Experiments on THINGS-EEG and THINGS-MEG under retrieval and generation protocols show consistent improvements over prior brain-signal decoding methods, while ablations confirm the complementary contributions of temporal factorization and CLIP-depth aggregation. These results suggest that effective brain-visual alignment

requires not only stronger brain encoders, but also a visual target whose embedding better matches the process and information distributed in brain responses.

Several directions remain for further refinement. Our experiments support the effectiveness of our core idea, and we will build upon this foundation to further refine the methodological details in future work. First, while the current decomposition strategy uses uniform temporal partitioning for brain signals and a fixed set of CLIP depths for visual features, this design already establishes a strong baseline. Future work could extend this to learned adaptive temporal boundaries or soft temporal assignments for brain signals, together with adaptive layer selection or depth weighting for CLIP features. Second, our current objective aligns the aggregated brain and visual representations through a global contrastive loss. A natural extension would be to introduce stage-wise or cross-depth supervision while preserving the global retrieval objective.

Acknowledgements

This work was partially supported by National Key R&D Program of China (No. 2025YFG0100400), National Distinguished Young Scholars (62325603) and CAS Project for Young Scientists in Basic Research (YSBR-116).

Disclosure of Interests

The authors declare no competing interests.

References

[1] Fernando Lopes da Silva. Eeg and meg: relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013. ISSN 0896-6273. doi: 10.1016/j.neuron.2013.10.017. URL https://doi.org/10.1016/j.neuron.2013.10.017.

[2] Nikos K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008. ISSN 1476-4687. doi: 10.1038/nature06976. URL https://doi.org/10.1038/nature06976.

[3] Sylvain Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, 20(3):327–339, 2017. ISSN 1546-1726. doi: 10.1038/nn.4504. URL https://doi.org/10.1038/nn.4504.

[4] Matteo Ferrante, Tommaso Boccato, and Nicola Toschi. Towards neural foundation models for vision: Aligning EEG, MEG and fMRI representations to perform decoding, encoding and modality conversion. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL https://openreview.net/forum?id=nxoKCdmteM.

[5] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

[6] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10760–10777, September 2023. ISSN 0162-8828. doi: 10.1109/TPAMI.2023.3263181. URL https://doi.org/10.1109/TPAMI.2023.3263181.

[7] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RxkcroC8qP.

[8] Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-mcrl: Neural multimodal contrastive representation learning for eeg-based visual decoding, 2025. URL https://arxiv.org/abs/2412.17337.

[9] Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. MB2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=5zAXR6HG4d.

[10] Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior, 2025. URL https://arxiv.org/abs/2503.04207.

[11] Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment, 2025. URL https://arxiv.org/abs/2511.06836.

[12] Xuhao Cao, Peiliang Gong, Liying Zhang, and Daoqiang Zhang. Eeg-clip: A transformer-based framework for eeg-guided image generation. *Neural Networks*, 194:108167, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108167. URL https://www.sciencedirect.com/science/article/pii/S0893608025010470.

[13] Chi-Sheng Chen. Necomimi: Neural-cognitive multimodal eeg-informed image generation with diffusion models, 2024. URL https://arxiv.org/abs/2410.00712.

[14] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, March 2014. ISSN 1546-1726. doi: 10.1038/nn.3635. URL https://doi.org/10.1038/nn.3635.

[15] Alex Clarke, Barry J. Devereux, Billi Randall, and Lorraine K. Tyler. Predicting the time course of individual objects with meg. *Cerebral Cortex*, 25(10):3602–3612, 10 2015. ISSN 1047-3211. doi: 10.1093/cercor/bhu203. URL https://doi.org/10.1093/cercor/bhu203.

[16] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, 2016. ISSN 2045-2322. doi: 10.1038/srep27755. URL https://doi.org/10.1038/srep27755.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[18] Jiawen Zheng, Haonan Jia, MING LI, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IEq71qS8B7.

[19] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J. Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq Mathew Abraham. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors, 2023. URL https://arxiv.org/abs/2305.18274.

[20] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, September 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-42891-8. URL https://doi.org/10.1038/s41598-023-42891-8.

[21] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

[22] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013. URL https://arxiv.org/abs/1311.2901.

[23] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.354. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.354.

[24] Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5023-14.2015. URL https://www.jneurosci.org/content/35/27/10005.

[25] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.

[26] Amin Ghiasi, Hamid Kazemi, Steven Reich, Eitan Borgnia, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration, 2023. URL https://openreview.net/forum?id=xc5ajsvLzF0.

[27] Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower, 2025. URL https://arxiv.org/abs/2504.21447.

[28] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks?, 2022. URL https://arxiv.org/abs/2108.08810.

[29] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlnC1rKPB.

[30] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, June 1996. ISSN 1476-4687. doi: 10.1038/381520a0. URL https://doi.org/10.1038/381520a0.

[31] Thomas Carlson, David A. Tovar, Arjen Alink, and Nikolaus Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of Vision*, 13(10):1, August 2013. ISSN 1534-7362. doi: 10.1167/13.10.1.

[32] K. Seeliger, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S.E. Bosch, and M.A.J. van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.07.018. URL https://www.sciencedirect.com/science/article/pii/S1053811917305864. New advances in encoding and decoding of brain signals.

[33] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.03.063. URL https://www.sciencedirect.com/science/article/pii/S1053811916300076.

[34] Yuelin Shi, Dasheng Bi, Janis K. Hesse, Frank F. Lanfranchi, Shi Chen, and Doris Y. Tsao. Rapid concerted switching of the neural code in the inferotemporal cortex. *Nature*, Mar 2026. doi: 10.1038/s41586-026-10267-3. URL https://doi.org/10.1038/s41586-026-10267-3.

[35] Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119754. URL https://www.sciencedirect.com/science/article/pii/S1053811922008758.

[36] Erin Gibson, Nancy J. Lobaugh, Steve Joordens, and Anthony R. McIntosh. Eeg variability: Task-driven or subject-driven signal of interest? *NeuroImage*, 252:119034, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119034. URL https://www.sciencedirect.com/science/article/pii/S105381192200163X.

[37] Zirui Lan, Olga Sourina, Lipo Wang, Reinhold Scherer, and Gernot R. Müller-Putz. Domain adaptation techniques for eeg-based emotion recognition: A comparative study on two public datasets. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):85–94, 2019. doi: 10.1109/TCDS.2018.2826840.

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

[41] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

[42] Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, feb 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580. URL https://doi.org/10.7554/eLife.82580.

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

[44] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information, 2024. URL https://arxiv.org/abs/2412.10489.

[45] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2024. URL https://arxiv.org/abs/2310.19812.

[46] Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from eeg signals, 2026. URL https://arxiv.org/abs/2602.22555.

[47] Jun-Mo Kim, Woohyeok Choi, Sang-Jun Park, Keun-Soo Heo, Young-Han Son, Ji-Hye Oh, Dong-Hee Shin, and Tae-Eui Kam. Seeeeg: Semantic-aware eeg-based multi-modal retrieval-augmented generation for high-fidelity visual brain decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4883–4892, October 2025.

[48] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion, 2023. URL https://arxiv.org/abs/2303.05334.

[49] Hongzhou Chen, Lianghua He, Yihang Liu, Longzhen Yang, Shaohua Shang, and MengChu Zhou. Visual neural decoding via improved visual-eeg semantic consistency, 2026. URL https://arxiv.org/abs/2408.06788.

[50] Chenyuan Hong, Jiaying Sun, Yifei Wu, Yao Jiang, Yueqin Diao, Lei Wang, and Huihui Fang. Lgfnet: A local-global focused network for eeg seizure detection. In *Proceedings of the 2025 6th International Conference on Intelligent Medicine and Health*, ICIMH '25, page 27–32, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400715747. doi: 10.1145/3784929.3784934. URL https://doi.org/10.1145/3784929.3784934.

[51] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

[52] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[54] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.08453.

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

A Additional dataset and implementation details

A.1 THINGS-EEG

THINGS-EEG [35, 42] is a large-scale EEG benchmark for object-centered visual decoding. At the dataset level, the released resource contains recordings from 10 participants and provides 82,160 trials per participant spanning 16,740 image conditions from the THINGS image database. The released data include both raw recordings and preprocessed derivatives. For comparability with recent EEG-based visual decoding methods, we adopt the benchmark split commonly used in prior work. Specifically, the training partition contains 1,654 object concepts, each associated with 10 images and

4 repetitions per image, whereas the test partition contains 200 held-out concepts, each associated with 1 image repeated 80 times. Consistent with prior EEG decoding studies Song et al. [5], we use the preprocessed signals and retain 17 posterior occipital-parietal channels [7, 10, 18] that are most relevant to visual processing: P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2. Repeated trials corresponding to the same visual stimulus are averaged within each subject to improve the signal-to-noise ratio and to obtain a single feature target per image during evaluation.

A.2 THINGS-MEG

THINGS-MEG [42] is a large-scale MEG benchmark from the THINGS-data collection. The full dataset contains recordings from 4 participants viewing images from 1,854 object concepts, with 12 exemplars per concept. Following recent MEG decoding protocols, we use the standard zero-shot split in which the training set contains 1,654 concepts with 12 images per concept and the test set contains 200 held-out concepts with 1 image per concept repeated 12 times. We use the released preprocessed MEG signals followed by a normalization for all experiments. As in the EEG setting, repeated trials of the same stimulus are averaged within each subject before downstream evaluation in order to improve signal quality.

A.3 Retrieval Evaluation protocol

For both datasets, we report zero-shot retrieval performance on a 200-way candidate set. When applicable, we evaluate both intra-subject and inter-subject settings. In the intra-subject setting, training and evaluation are performed separately for each participant. In the inter-subject setting, the model is trained on all remaining participants and evaluated on the held-out participant following the standard leave-one-subject-out protocol. Unless otherwise stated, all reported results are based on the official benchmark partitions and the repetition-averaged test representations described above.

A.4 Experimental settings

Unless otherwise specified, for retrieval, the model is trained with AdamW, using initial learning rate $5e^{-4}$ under a cosine decay schedule with a weight decay of `0.05` and an initial temperature of `0.07`. We train the model for `25` epochs with a batch size of `1024`. The shared embedding dimension d and the hidden dimension s are set to `1024`. While the number of temporal segments in the brain encoder is set to `2`, the visual encoder is LAION CLIP ViT-B/32 [17, 43] by default, along with the selected depth set is $|\mathcal{L}| = 5$ (4 intermediate layers plus the final embedding). For generation, the prior is pretrained for `100k` steps with learning rate `1e-4` and a batch size of `4`. At inference time, we use SDXL-Turbo with a `4`-step sampler [51]. All experiments are conducted on a single NVIDIA 4090 24 GB GPU. Most experiments can be completed within a short running time, while the prior pretraining stage takes approximately 25 hours.

B Why this multiscale design

The proposed design is intentionally asymmetric. Brain signals exhibit rich temporal dynamics, making temporal partitioning and segment aggregation natural modeling choices. In contrast, the visual modality already admits a strong pretrained hierarchy, so its multiscale structure is more naturally captured by features drawn from multiple depths of a frozen CLIP encoder. The resulting shared space is therefore supervised by a visual target that is richer than a single semantic vector, as brain representations do. The model remains lightweight, parameter-efficient, and explicitly interpretable at the segment level.

C Baseline subtraction

Brain signals often exhibit substantial subject-dependent static offsets, especially in inter-subject settings where cross-subject variability can dominate task-relevant structure [36, 37]. To reduce this nuisance variation, we optionally apply **baseline subtraction** before the temporal partition and segment encoding stages. Specifically, for each subject s, we estimate a subject-specific prototype

with an unlabeled calibration set:

$$\mu_s = \frac{1}{|\mathcal{C}_s|} \sum_{X_j \in \mathcal{C}_s} X_j, \tag{17}$$

where $\mathcal{C}_s$ denotes a calibration set associated with subject s, $|\mathcal{C}_s|$ is the number of calibration trials for that subject, and $\mu_s \in \mathbb{R}^{C \times T}$ represents the subject-level baseline pattern. We then subtract this prototype from each trial:

$$\mathcal{X}_i = X_i - \mu_{s_i}, \tag{18}$$

We apply baseline subtraction only in the *inter-subject* setting, where explicit calibration helps suppress subject-specific offsets and improves cross-subject alignment. In the *intra-subject* setting, where training and evaluation already take place within the same subject and such between-subject variability is not the primary challenge, we do not apply baseline subtraction and instead retain the original signals.

D Generation pipeline

D.1 Related work

Recent advances in generative modeling have substantially improved visual reconstruction from neural signals. Early brain-to-image generation studies mainly rely on aligning brain representations with pretrained visual or multimodal latent spaces, such as VAE latents [52], as compact visual targets for image synthesis, and then using generative priors to synthesize images from the decoded representation. A key challenge in brain-to-image generation is how to bridge the distribution gap between noisy brain-derived embeddings and the conditioning space expected by pretrained image generators. Latent diffusion models [53] and SDXL [40] provide strong image priors, but directly conditioning them on neural features can lead to unstable or semantically misaligned reconstructions. Adapter-based diffusion methods offer an effective solution by introducing lightweight conditioning modules while keeping the generative backbone largely frozen. T2I-Adapter [54] and ControlNet [55] inject external structural conditions into pretrained diffusion models, while IP-Adapter [41] decouples image-prompt conditioning from text conditioning through additional cross-attention layers. Following this line of work, BrainHiVE [18] adopts a generation-side architecture that first learns a visual-to-diffusion prior and then aligns brain representations to this pretrained visual conditioning space. This design separates prior learning from brain-side alignment, reducing the burden of learning a direct brain-to-image mapping from limited and noisy neural data.

D.2 Generation architecture

For brain-to-image generation, we follow the generation-side design of BrainHiVE [18]. The generation branch consists of a Brain Granularity Encoder (BGE), a Visual Depth Encoder (VDE), a lightweight conditioning projector, an IP-Adapter and a frozen diffusion generator. The BGE encodes brain signals into a generation-oriented latent representation, while the VDE encodes the paired visual stimulus into the visual target space used for diffusion conditioning.

Given an image x_v, the VDE produces a visual representation

$$z_v = f_{\mathrm{VDE}}(x_v). \tag{19}$$

This representation is mapped into the diffusion conditioning space by a residual projector:

$$z_c = z_v + \phi_c(z_v), \tag{20}$$

where ϕ_c is a two-layer MLP with GELU activation. The projected condition z_c is injected into a frozen SDXL UNet [40] through an IP-Adapter [41], following the adapter-based conditioning mechanism adopted by BrainHiVE [18]. The diffusion backbone is kept frozen throughout prior training and brain-side alignment.

D.3 Prior pretraining

We first pretrain the vision-to-diffusion prior using image data only. During this stage, the image is encoded by VDE, projected into the conditioning space, and used to guide the frozen diffusion model

Table 5: Subject-wise Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG under the intra-subject setting.

Method	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
***Intra-Subject**: train and test on one subject*																						
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
ATM	25.6	60.4	22.0	54.5	25.0	62.4	31.4	60.9	12.9	43.0	21.3	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	27.1	58.1
CogCapturer	31.4	79.7	31.4	77.8	38.2	85.7	40.4	85.8	24.4	66.3	34.8	78.8	34.7	81.0	48.1	88.6	31.4	79.4	35.6	79.3	35.6	80.2
Neural-MCRL	27.5	64.0	28.5	61.5	37.0	69.0	35.0	66.0	22.5	51.5	31.5	61.0	31.5	62.5	42.0	74.5	30.5	59.5	37.5	71.0	32.4	64.1
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	70.0
UBP	41.2	70.5	51.2	80.9	51.2	82.0	51.1	76.9	42.2	72.8	57.5	83.5	49.0	79.9	58.6	85.8	45.1	76.2	61.5	88.2	50.9	79.7
SRT	23.0	55.1	23.0	57.3	28.0	59.0	31.5	64.4	20.6	49.8	21.4	58.9	26.0	60.5	36.2	66.5	27.6	61.6	26.0	62.1	26.3	59.5
NeuroBridge	50.0	77.6	63.2	90.6	61.6	91.1	61.4	90.0	54.8	85.0	69.7	92.9	62.7	88.8	71.2	95.1	64.0	91.0	73.6	97.1	63.2	89.9
AVDE	25.0	55.2	24.1	51.0	27.5	58.6	29.8	54.7	25.4	50.3	33.5	60.3	27.4	55.2	41.7	71.3	26.1	52.1	39.5	73.0	30.0	58.2
BrainHiVE	64.3	88.8	76.3	95.3	74.0	95.0	67.0	91.8	68.0	91.5	81.5	96.3	76.8	**96.8**	84.8	98.5	76.8	95.8	**87.3**	99.3	75.7	94.6
BRIDGE(Ours)	**77.5**	**93.0**	**80.5**	**98.2**	**80.8**	**97.0**	**79.3**	**97.3**	**72.7**	**94.5**	**85.0**	**99.5**	**80.2**	95.8	**88.0**	**99.0**	**77.7**	**96.3**	85.8	**99.7**	**80.8**	**97.0**

through IP-Adapter. Given a noisy latent x_t at timestep t, the model is optimized with the standard diffusion noise-prediction objective:

$$\mathcal{L}_{\text{prior}} = \|\epsilon - \delta(x_t, t, z_c)\|_2^2 , \tag{21}$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes the sampled Gaussian noise and $\delta(\cdot)$ denotes the frozen denoising UNet conditioned on z_c. Text prompts are kept empty so that the model learns an image-conditioned, text-free mapping from VDE features to the diffusion conditioning space.

D.4 Brain-to-generation alignment

After prior pretraining, we freeze the VDE, conditioning projector, IP-Adapter, and SDXL UNet. Only the BGE is trained on paired brain-image samples. For each pair (x_b, x_v), BGE produces the brain-side generation representation

$$z_b = f_{\text{BGE}}(x_b), \tag{22}$$

which is aligned with the VDE representation z_v though contrastive learning. This alignment forces the brain-derived embedding to match the visual representation distribution expected by the pretrained generation prior, reducing the instability caused by directly conditioning the diffusion model on noisy neural features.

At inference time, only the brain signal x_b is required. We compute $z_b = f_{\text{BGE}}(x_b)$ and feed it into the pretrained conditioning pathway to obtain the diffusion condition. The frozen IP-Adapter and SDXL UNet then synthesize the reconstructed image. Following BrainHiVE [18], this design separates visual-to-diffusion prior learning from brain-to-visual alignment, enabling stable generation from brain signals.

E Training procedure

The training procedure is as shown in Algorithm 1.

F Results details

F.1 Retrieval results

We report detailed subject-wise retrieval results for both THINGS-EEG and THINGS-MEG in Tables 5–8. Additionally, we provide subject-wise ablation results in Tables 9–12, covering the contributions of the BGE, VDE, and the choices of EEG and image stages.

F.2 Generation results

BRIDGE is implemented by building upon the BrainHiVE [18] framework and following the same parameter settings, with details provided in Appendix A. In Table 13, BrainHiVE† denotes the results reproduced in our experimental environment. We further compare BRIDGE with representative state-of-the-art methods. The results show that BRIDGE achieves state-of-the-art performance too, while different methods exhibit complementary strengths across datasets and metrics.

Algorithm 1 Brain-visual multiscale representation alignment training

Require: Batch $\{(\mathcal{X}_i, \mathcal{I}_i, s_i)\}_{i=1}^{B}$, segment count S, CLIP depth set $\mathcal{L}$

for $i = 1$ to B **do**
  if baseline subtraction enabled **then**
    $\widetilde{\mathcal{X}}_i \leftarrow \mathcal{X}_i - \mu_{s_i}$ ▷ μ_{s_i} computed from non-evaluation calibration/training data
  else
    $\widetilde{\mathcal{X}}_i \leftarrow \mathcal{X}_i$
  end if
  Uniformly partition $\widetilde{\mathcal{X}}_i$ into S temporal segments $\{\widetilde{\mathcal{X}}_i^{(m)}\}_{m=1}^{S}$
  for $m = 1$ to S **do**
    Zero-pad $\widetilde{\mathcal{X}}_i^{(m)}$ along the temporal axis to length $T_{\max}$
    $u_i^{(m)} \leftarrow \mathrm{Flatten}\Big(\mathrm{Pad}\Big(\widetilde{\mathcal{X}}_i^{(m)}\Big)\Big)$
    $\bar{h}_i^{(m)} \leftarrow u_i^{(m)} W_b$
    $h_i^{(m)} \leftarrow \mathrm{LayerNorm}\Big(\bar{h}_i^{(m)} + \phi_b\Big(\bar{h}_i^{(m)}\Big)\Big)$
  end for
  for $m = 1$ to S **do**
    $a_i^{(m)} \leftarrow \rho_\omega\Big(h_i^{(m)}\Big)$
  end for
  $\alpha_i^{(m)} \leftarrow \dfrac{\exp(a_i^{(m)})}{\sum_{r=1}^{S} \exp(a_i^{(r)})}, \quad m = 1, \ldots, S$
  $z_i^b \leftarrow \sum_{m=1}^{S} \alpha_i^{(m)} h_i^{(m)}$
  Extract frozen CLIP features $\{v_i^{(\ell)}\}_{\ell \in \mathcal{L}}$ from $\mathcal{I}_i$
  for each $\ell \in \mathcal{L}$ **do**
    $b_i^{(\ell)} \leftarrow W_\ell v_i^{(\ell)} + c_\ell$
  end for
  $b_i \leftarrow \mathrm{Concat}\big(b_i^{(\ell)}\big)_{\ell \in \mathcal{L}}$
  $\bar{z}_i^v \leftarrow b_i W_v$
  $z_i^v \leftarrow \mathrm{LayerNorm}(\bar{z}_i^v + \phi_v(\bar{z}_i^v))$
end for
L2-normalize $\{z_i^b\}_{i=1}^{B}$ and $\{z_i^v\}_{i=1}^{B}$ to obtain $\{\hat{z}_i^b\}_{i=1}^{B}$ and $\{\hat{z}_i^v\}_{i=1}^{B}$
Form the similarity matrix using Eq. (15)
Compute the symmetric contrastive loss using Eq. (16)
Update all trainable parameters with gradient descent

Table 6: Subject-wise Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG under the inter-subject setting.

Method	Sub 1		Sub 2		Sub 3		Sub 4		Sub 5		Sub 6		Sub 7		Sub 8		Sub 9		Sub 10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
***Inter-Subject**: leave one subject out for test*																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	6.2	21.4
ATM	10.5	26.8	7.1	24.8	11.9	33.8	14.7	39.4	7.0	23.9	11.1	35.8	16.1	43.5	15.0	40.3	4.9	22.7	20.5	46.5	11.9	33.8
Neural-MCRL	13.0	31.5	12.0	30.5	14.5	35.5	12.5	35.5	11.5	29.0	13.5	35.5	14.0	36.0	18.5	38.5	13.5	32.5	17.0	39.0	14.0	34.3
UBP	11.5	29.7	15.5	40.0	9.8	27.0	13.0	32.3	8.8	33.8	11.7	31.0	10.2	23.8	12.2	32.2	15.5	40.5	16.0	43.5	12.4	33.4
SRT	15.2	37.5	16.7	37.8	11.9	35.9	13.1	37.3	13.1	33.3	14.6	39.1	8.8	28.5	12.1	34.1	11.1	34.7	21.1	52.5	13.8	37.1
NeuroBridge	23.2	52.4	21.2	49.3	13.2	36.5	17.0	45.3	14.5	37.7	25.0	55.0	15.3	45.1	20.1	44.9	13.7	36.5	27.2	56.3	19.0	45.9
AVDE	14.1	32.2	17.0	38.4	9.1	21.8	15.2	32.5	12.5	32.4	17.3	38.6	7.4	20.4	18.5	40.1	13.2	33.6	18.0	39.3	14.3	32.9
BrainHiVE	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	20.0	44.1
BRIDGE(Ours)	**40.2**	**72.7**	**42.5**	**74.7**	**26.7**	**56.5**	**30.0**	**57.2**	**27.7**	**57.0**	**40.7**	**74.5**	**28.8**	**64.5**	**28.2**	**58.0**	**30.0**	**64.2**	**42.2**	**79.0**	**33.7**	**65.8**

On EEG, SeeEEG [47] performs best on several high-level semantic metrics, whereas BRIDGE achieves the best PixCorr. In the EEG subject-8 setting, AVDE [46] achieves the best SSIM, Inception, CLIP, and SwAV scores, while BRIDGE obtains the best AlexNet(2) score, and remains competitive on PixCorr and AlexNet(5). On MEG, existing methods perform strongly on different aspects of reconstruction quality. For example, AVDE achieves the best PixCorr and Inception scores, B.D. obtains the best SSIM, while BRIDGE achieves competitive semantic reconstruction performance, obtaining the best SwAV score and strong results on other high-level metrics.

Table 7: Subject-wise Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG under the intra-subject setting.

Method	Sub1		Sub2		Sub3		Sub4		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
***Intra-Subject**: train and test on one subject*										
NICE	9.6	27.8	18.5	47.8	14.2	41.6	9.0	26.6	12.8	36.0
NICE-SA	9.8	27.8	18.6	46.4	10.5	38.4	11.7	27.2	12.7	35.0
NICE-GA	8.7	30.5	21.8	56.6	16.5	49.7	10.3	32.3	14.3	42.3
UBP	15.0	38.0	46.0	80.5	27.3	59.0	18.5	43.5	26.7	55.2
AVDE	22.1	49.8	21.1	46.6	24.4	53.3	26.6	49.4	23.6	49.8
BrainHiVE	**14.0**	31.8	**63.8**	**91.8**	41.0	**78.3**	17.0	41.0	**33.9**	60.7
BRIDGE(Ours)	11.2	**33.5**	55.0	88.2	**44.8**	76.8	**21.5**	**47.3**	33.1	**61.5**

Table 8: Subject-wise Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-MEG under the inter-subject setting.

Method	Sub1		Sub2		Sub3		Sub4		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
***Inter-Subject**: leave one subject out for test*										
UBP	2.0	5.7	1.5	17.2	2.7	10.5	2.5	8.0	2.2	10.4
BrainHiVE	–	–	–	–	–	–	–	–	5.4	15.2
BRIDGE(Ours)	**5.3**	**12.7**	**9.2**	**22.0**	**8.2**	**23.8**	**6.8**	**18.5**	**7.4**	**19.3**

Overall, these results indicate that BRIDGE reaches state-of-the-art reconstruction quality under a consistent and reproducible evaluation protocol. Figure 3 shows randomly selected generated image samples. Subject-wise reconstruction metrics are summarized in Table 14 and Table 15.

G LDA visualization

We further provide clustering visualizations in Fig. 4–Fig. 5 as additional evidence that splitting the EEG into two temporal segments is the most reasonable and robust design choice. In the 1D LDA projection, the two-stage setting yields a clean separation between the two classes under a simple linear discriminant, indicating that the corresponding temporal features are highly stable and readily distinguishable. By contrast, when the EEG is partitioned into three, four, five, or six segments, the projected distributions exhibit substantial overlap, showing that these finer-grained segments are less consistently separable and therefore more sensitive to noise and sampling variation. The 2D LDA plots for the multi-stage settings further support this interpretation. Although some local structure remains visible, the clusters are no longer cleanly isolated and instead interfere with one another, which suggests weaker discriminability and less reliable cross-modal alignment. More broadly, these figures reveal a › trend that is consistent with the quantitative results, the more distinctly the segment-level features are separated, the higher the downstream retrieval performance tends to be.

Table 9: Subject-wise joint ablation results. Results are reported as Top-1 and Top-5 accuracies (%).

Setting	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
w/o BGE	73.83	91.33	75.83	98.00	78.67	96.67	76.83	96.83	66.50	92.50	83.00	97.83	73.83	95.00	86.00	99.33	71.83	94.67	85.00	98.83	77.10	96.10
w/o VDE	41.00	75.00	48.00	81.17	52.17	84.67	55.50	84.83	43.83	76.17	55.00	83.50	51.67	82.50	59.67	85.50	46.17	80.00	65.50	93.83	51.85	82.72
w/o both	38.50	74.17	44.50	76.67	48.67	81.67	51.50	84.83	40.67	72.67	51.33	82.17	47.67	78.50	60.83	88.00	43.17	76.67	63.00	93.67	48.98	80.90
FULL	**77.50**	**93.00**	**80.50**	**98.17**	**80.83**	**97.00**	**79.33**	**97.33**	**72.67**	**94.50**	**85.00**	**99.50**	**80.17**	**95.83**	**88.00**	**99.00**	**77.67**	**96.33**	**85.83**	**99.67**	**80.75**	**97.03**

Table 10: Subject-wise ablation study on the number of EEG stages. Results are reported as Top-1 and Top-5 accuracies (%).

EEG stages S	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
2	**71.83**	**91.67**	**78.17**	**96.50**	**78.17**	**96.00**	**71.33**	**93.00**	**74.17**	**94.67**	**82.17**	**99.00**	**75.33**	**95.50**	**86.33**	**98.67**	**81.83**	**96.83**	**89.33**	**99.67**	**78.87**	**96.15**
3	64.00	87.00	74.50	96.33	75.50	95.17	69.17	90.67	72.33	93.17	79.50	97.50	75.17	93.17	85.50	98.17	77.17	95.67	86.33	99.00	75.92	94.58
4	56.33	83.33	64.00	91.83	65.50	92.33	56.00	84.67	58.50	87.00	68.17	94.33	67.83	91.17	82.33	97.17	66.50	93.50	78.17	98.00	66.33	91.33
5	31.83	64.17	40.00	71.33	32.17	60.00	24.83	53.17	17.33	37.33	11.33	23.83	37.50	63.50	60.17	88.17	22.67	54.83	57.33	90.00	33.52	60.63
6	36.83	67.67	40.17	73.33	48.83	78.33	36.67	68.67	33.50	69.33	48.33	79.50	48.67	81.50	65.17	89.17	48.17	79.17	60.17	90.00	46.65	77.67

Table 11: Subject-wise ablation study on the number of image stages. Results are reported as Top-1 and Top-5 accuracies (%).

Image stages $\lvert\mathcal{L}\rvert$	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
1 (5)	69.50	91.00	70.67	94.17	73.83	94.17	73.50	95.17	61.00	89.67	79.33	98.00	74.83	93.33	82.83	99.00	70.00	93.17	83.33	99.00	73.88	94.67
2 (5, 11)	69.33	88.67	71.50	95.17	75.00	94.83	72.17	94.83	63.17	90.50	78.50	97.50	71.67	92.83	82.17	98.50	68.67	92.83	80.17	99.33	73.23	94.50
3 (3, 7, 11)	72.50	90.00	73.83	95.67	76.17	96.33	76.17	95.17	66.17	91.50	80.33	97.67	73.67	94.50	84.67	99.67	69.17	93.67	84.67	99.00	75.73	95.32
4 (2, 5, 8, 11)	73.83	91.33	75.83	98.00	78.67	96.67	76.83	96.83	66.50	92.50	83.00	97.83	73.83	95.00	86.00	99.33	71.83	94.67	85.00	98.83	77.13	96.10
5 (2, 4, 6, 8, 11)	73.33	90.67	77.17	96.50	79.17	96.83	78.33	97.17	67.50	92.50	83.67	98.50	73.17	95.33	87.33	99.33	74.33	94.67	84.50	99.50	77.85	96.10
6 (1, 3, 5, 7, 9, 11)	73.33	92.17	77.50	97.17	82.17	96.50	78.17	97.50	71.17	93.83	83.17	98.33	76.33	95.17	86.33	99.83	76.67	96.17	86.50	99.83	79.10	96.65

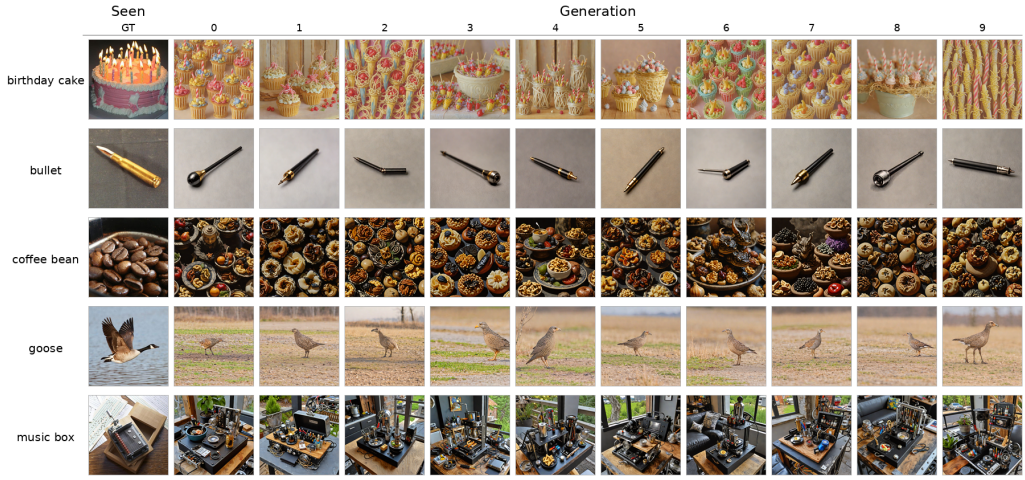
Table 12: Subject-wise joint stage ablation results. Results are reported as Top-1 and Top-5 accuracies (%).

EEG	Image	Sub1		Sub2		Sub3		Sub4		Sub5		Sub6		Sub7		Sub8		Sub9		Sub10		Avg	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
2	6	77.66	93.83	83.33	97.33	81.33	97.83	81.50	97.50	74.67	93.50	87.00	99.33	81.67	95.00	88.67	99.50	82.33	97.00	87.67	99.83	82.58	97.06
4	4	61.50	84.50	72.00	96.17	66.67	93.67	59.83	91.00	65.33	91.00	75.17	96.00	69.83	91.83	80.83	97.33	63.50	91.00	78.17	96.33	69.28	92.88
6	4	47.00	76.17	47.33	79.67	45.67	83.17	40.83	73.67	44.83	76.50	46.50	75.67	57.50	84.67	56.17	89.50	49.50	83.17	54.33	86.67	48.97	80.88
2	2	73.83	90.83	76.83	96.17	76.17	96.67	75.17	95.83	70.83	94.33	82.67	99.17	75.67	94.00	86.83	99.00	76.17	95.33	84.33	98.67	77.85	96.00
4	2	58.00	83.83	66.33	92.33	67.00	93.17	60.50	88.83	59.17	88.33	72.33	95.50	67.17	89.33	81.67	97.50	60.50	88.83	73.83	95.33	66.65	91.29

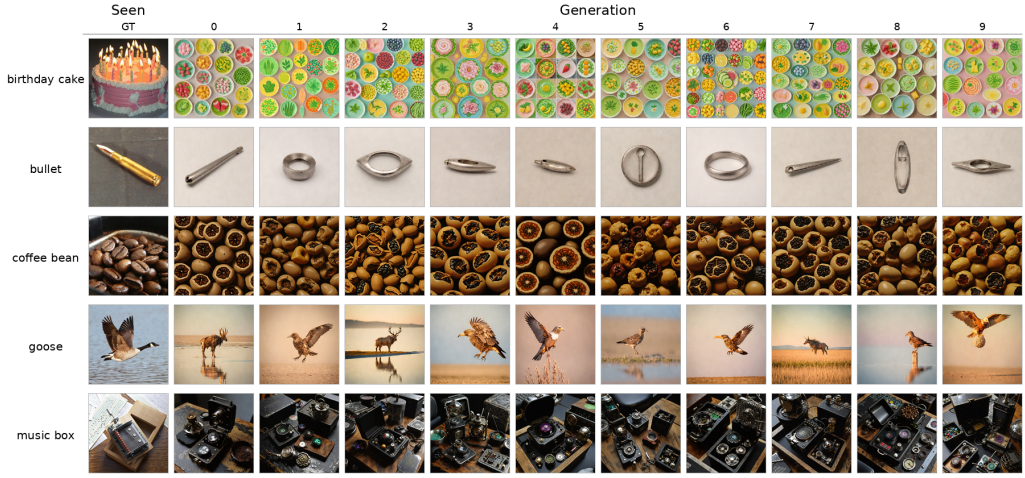
Table 13: Quantitative assessments of the reconstruction quality for EEG and MEG.

Dataset	Method	Low-level		High-level				
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
MEG	B.D. [45]	0.076	0.336	0.736	0.826	0.671	0.767	0.584
	ATM [7]	0.104	0.340	0.613	0.672	0.619	0.603	0.651
	AVDE [46]	0.124	0.339	0.728	0.801	0.685	0.720	0.651
	BrainHiVE† [18]	0.111	0.261	0.670	0.763	0.658	0.695	0.646
	BRIDGE (Ours)	0.104	0.258	0.672	0.774	0.679	0.697	0.644
EEG	C.C.(All) [44]	0.150	0.347	0.754	0.623	0.669	0.715	0.590
	SeeEEG [47]	0.130	0.344	0.809	0.904	0.784	0.869	0.539
	BrainHiVE† [18]	0.162	0.298	0.767	0.827	0.671	0.725	0.610
	BRIDGE (Ours)	0.163	0.304	0.799	0.855	0.699	0.736	0.602
EEG (subj-8)	ATM [7]	0.160	0.345	0.776	0.866	0.734	0.786	0.582
	AVDE [46]	0.188	0.396	0.817	0.889	0.765	0.795	0.557
	BrainHiVE† [18]	0.189	0.316	0.788	0.836	0.682	0.746	0.601
	BRIDGE (Ours)	0.188	0.332	0.821	0.868	0.726	0.764	0.584

† Results reproduced by us. The originally reported metrics could not be reproduced exactly under our experimental settings.



(a) Subject 8



(b) Subject 10

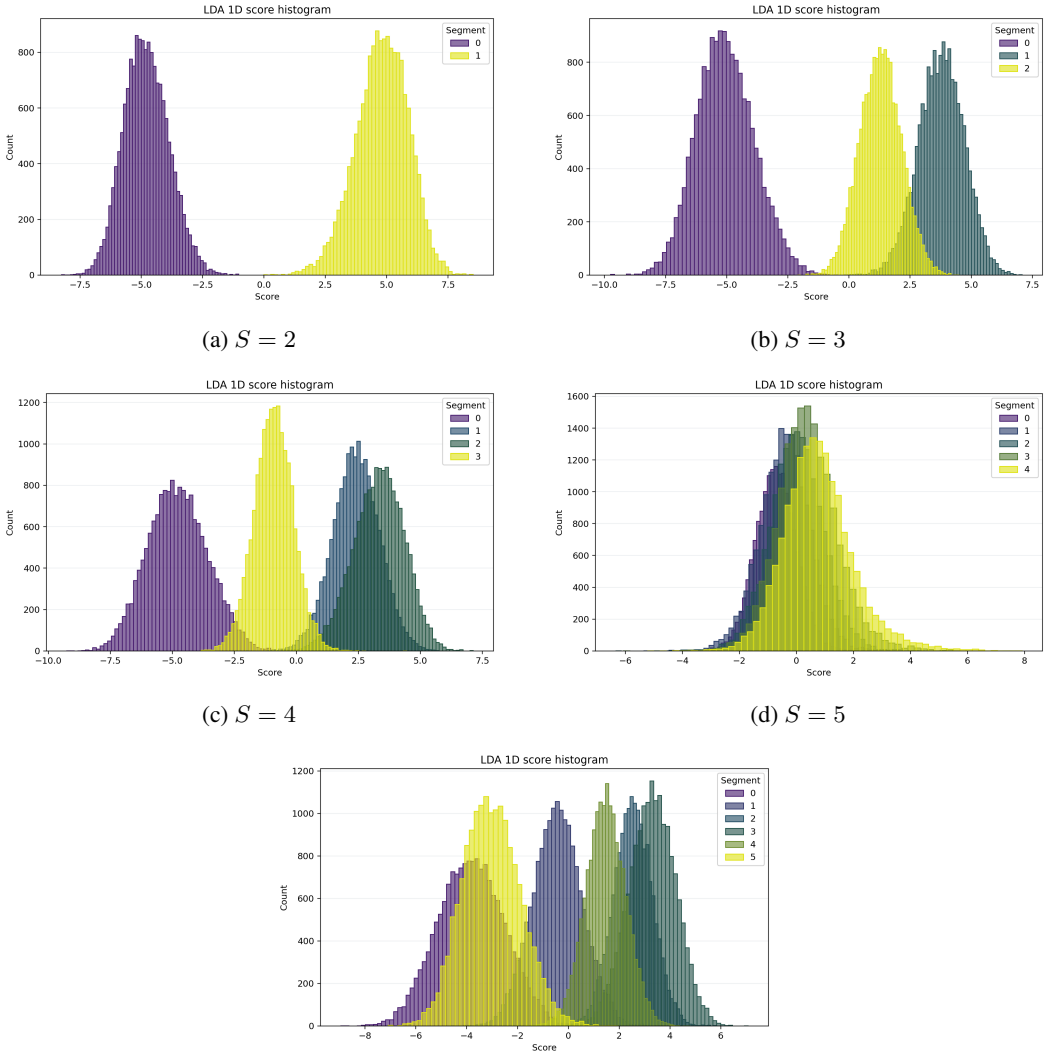
Figure 3: Randomly selected generated image samples on THINGS-EEG for subject 8 and subject 10. The first column shows the seen image, and the remaining columns show ten generated samples.

Table 14: BRIDGE reconstruction results across subjects on THINGS-EEG.

Subject	**Low-level**		**High-level**				
	Pixcorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
1	0.158	0.282	0.795	0.830	0.672	0.684	0.636
2	0.165	0.317	0.826	0.873	0.693	0.751	0.599
3	0.164	0.305	0.801	0.867	0.716	0.754	0.598
4	0.166	0.298	0.788	0.857	0.703	0.751	0.598
5	0.151	0.284	0.789	0.847	0.691	0.715	0.608
6	0.174	0.293	0.803	0.855	0.695	0.734	0.610
7	0.143	0.304	0.779	0.843	0.685	0.726	0.613
8	0.188	0.332	0.821	0.868	0.726	0.764	0.584
9	0.148	0.295	0.783	0.846	0.695	0.716	0.589
10	0.168	0.328	0.800	0.867	0.712	0.770	0.586
Ave	**0.163**	**0.304**	**0.799**	**0.855**	**0.699**	**0.736**	**0.602**

Table 15: BRIDGE reconstruction results across subjects on THINGS-MEG.

Subject	Low-level		High-level				
	Pixcorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	SwAV ↓
1	0.086	0.215	0.608	0.719	0.642	0.668	0.653
2	0.110	0.293	0.724	0.831	0.732	0.752	0.619
3	0.106	0.281	0.711	0.802	0.708	0.715	0.634
4	0.114	0.243	0.645	0.745	0.632	0.651	0.670
Ave	**0.104**	**0.258**	**0.672**	**0.774**	**0.679**	**0.697**	**0.644**

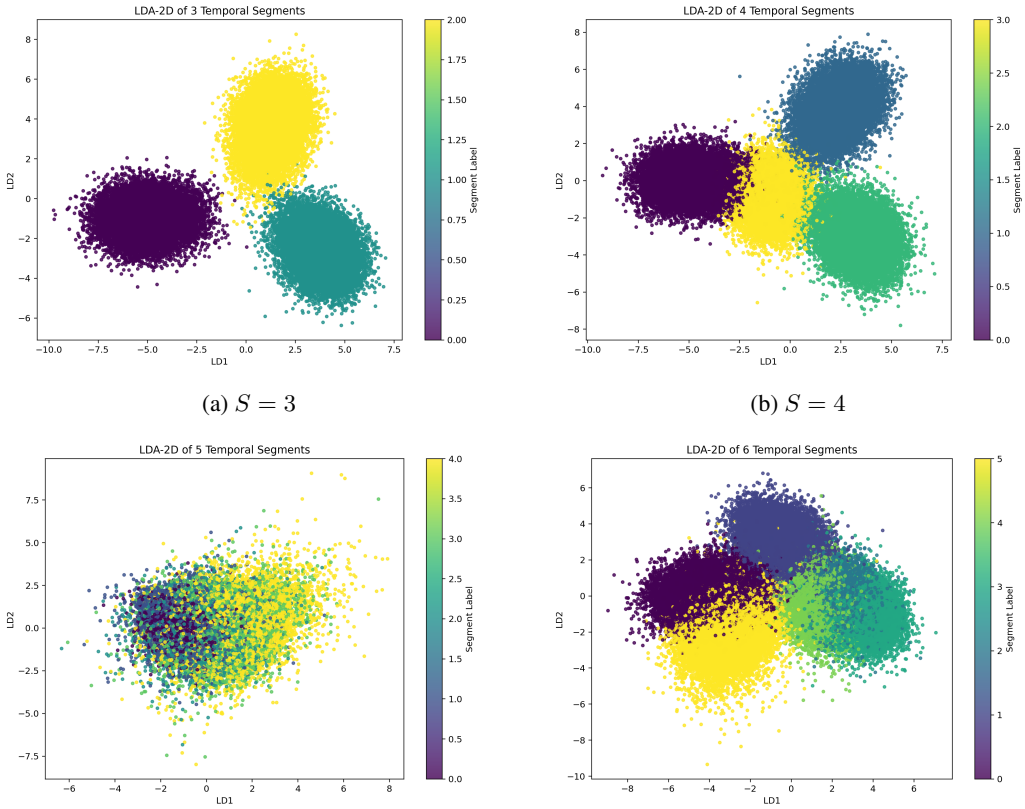


(a) $S = 2$ (b) $S = 3$

(c) $S = 4$ (d) $S = 5$

(e) $S = 6$

Figure 4: LDA 1D score histograms for five segment settings of subject-01. Subfigures (a)–(e) correspond to different numbers of segments. The distributions exhibit visibly separated patterns across segment settings, suggesting that the extracted features are reasonably separable.



(a) $S = 3$ (b) $S = 4$

(c) $S = 5$ (d) $S = 6$

Figure 5: LDA 2D score maps for four segment settings of subject-01. Subfigures (a)–(d) correspond to different numbers of segments. The distributions exhibit visibly separated patterns across segment settings, suggesting that the extracted features are reasonably separable.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction first identify a fundamental hierarchical representation mismatch in existing brain-signal-based visual decoding methods. Based on this observation, the paper proposes BRIDGE, which addresses this mismatch by aligning brain signals and visual representations through Brain Granularity Encoding and Visual Depth Encoding. The claimed contributions and improvements are supported by retrieval, reconstruction, and ablation experiments on THINGS-EEG and THINGS-MEG.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We dedicate a paragraph in Section 5 to explain the limitations of our work and future improvement directions in details.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is primarily an empirical method paper and does not introduce formal theoretical results, theorems, or proofs. The mathematical formulations are used to define the proposed model components and training objective.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper method specifies the main architecture of BRIDGE, and the generation pipeline, together with additional details presented in Appendix A. The code will further provide implementation-level details needed to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code, datasets, model implementation, and detailed reproduction instructions, are included in the supplementary material and will be released on GitHub upon acceptance to enable faithful reproduction of the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the configuration of the training and test details in Experiments section and additional details are provided in Appendix A and code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report subject-wise results and cross-subject averages, with three repeated runs using different random seeds for the main retrieval and reconstruction experiments. These results characterize variability across both subjects and runs, while the ablation studies further support the robustness of the proposed BGE and VDE components.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the information of computer resources in detail in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses publicly available EEG/MEG visual decoding benchmarks and does not involve new data collection from human participants. The research fully follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper studies non-invasive brain-signal-based visual decoding, which may benefit neuroscience and cognitive modelin. The method uses authorized public datasets. Therefore, we do not identify additional negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release a new high-risk dataset or a large-scale generative foundation model.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All of the code, datasets and models we utilize are explicitly cited in the paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include the code and datasets details of our method in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We use publicly available datasets from human subjects studies and do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: We use publicly available datasets from human subjects studies and do not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We only used LLMs for grammar checking and language polishing to improve readability.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.